CIENCIA DE LOS DATOS

Detectives del código genético

En la era del *big data*, un equipo de investigadores del INTA aplica herramientas de la bioinformática para ordenar, secuenciar y analizar grandes volúmenes de datos. Descifrar el ADN de un organismo vivo permite entender cómo funciona y cuáles son los mecanismos que se activan frente a una enfermedad o cambios en el ambiente.

POR CECILIE ESPERBENT
FOTOGRAFÍA MATÍAS OTTAVIANI

Conformado por dos hebras de ADN enrolladas en forma de hélice, que dan origen a cada uno de los 23 pares de cromosomas (cada cromosoma tiene entre 50.000.000 y 300.000.000 de pares de bases), el tamaño del genoma humano es de 32.000 millones de bases. Por el gran caudal de datos que implica su análisis, haber descifrado esa secuencia fue uno de los mayores logros biomédicos de los últimos años. Conocer el orden exacto de los pares de bases en un segmento de ADN permitirá, en el futuro, descifrar mecanismos que luego podrán ayudar a paliar o evitar enfermedades.

En 2003, la secuenciación del genoma humano revolucionó la manera de abordar el estudio del ADN. Su ordenamiento fue posible gracias a los avances en métodos usados para analizar ácidos nucleicos y al desarrollo de tecnologías cada vez más sofisticadas de secuenciación. Además, la bioinformática facilitó el análisis masivo de datos y su integración con conocimientos previos aportados por años de estudios de genética humana.

> "La bioinformática responde a las necesidades de procesamiento, almacenamiento y análisis de datos biológicos para generar nueva información y conocimientos" (Maximo Rivarola).

A pesar de los múltiples progresos en biología e informática, secuenciar todo el ADN de un organismo sigue siendo una tarea compleja. Sin embargo, gracias a nuevos métodos, ahora ordenar un genoma es mucho más rápido y menos costoso de lo que resultó en el Proyecto Genoma Humano.

Con el transcurrir de los años, los logros de la genética molecular y poblacional, sumado a la biología celular fueron acompañados de los avances computacionales necesarios para el procesamiento de la información genética, desde algoritmos o modelos computacionales capaces de responder preguntas relacionadas con la variación en las secuencias de los genes, hasta el desarrollo de equipos con la capacidad para almacenar la información y consultarla eficientemente.

En la actualidad, resulta sencillo imaginarnos el trabajo en un laboratorio vinculado con las computadoras, pero esto no siempre fue así. De hecho, antes de 1990 no se conocía la secuencia del genoma de ningún organismo. Recién en 1995 se publicaron los códigos genéticos de las bacterias *Haemophilus influenzae* y *Mycoplasma genitalium*.

A 20 kilómetros de la Ciudad Autónoma de Buenos Aires, en la localidad de Hurlingham, funciona el Centro Nacional de Investigaciones Agropecuarias –CNIA– del INTA. Pocos saben a qué se dedican las más de 1.300 personas que trabajan en cuatro centros de investigación –divididos en 16 institutos–. Sin embargo, allí se concentra gran parte del trabajo científico que realiza el organismo.

En el marco del Centro de Investigación en Ciencias Veterinarias y Agronómicas –CICVyA–, funciona la Unidad de Bioinformática. En ese lugar, técnicos especializados e investigadores trabajan en red con pares de distintas unidades del INTA para desentrañar la

información genética de especies forestales, frutales, cereales y oleaginosas, plagas, malezas y patógenos. Son detectives que buscan entender la arquitectura genética de organismos de interés agrícola.

Biólogos, matemáticos, técnicos de laboratorio y bioinformáticos articulan sus tareas diarias en busca de respuestas a estudios exhaustivos sobre un problema biológico determinado. En todos los casos, generan una gran cantidad de datos que demandan soluciones bioinformáticas, tanto para su ordenamiento como para su análisis.

Máximo Rivarola es biólogo molecular y trabaja en investigaciones vinculadas con el procesamiento masivo de datos de genómica en el ámbito de la agrobiotecnología. Como referente en bioinformática del INTA, integró consorcios internacionales para la secuenciación del genoma del trigo, girasol y bacterias de interés agrícola.

"La bioinformática es una disciplina que ha evolucionado rápidamente", señaló Rivarola y agregó: "Responde al avance y a las necesidades de procesamiento, almacenamiento y análisis de datos biológicos derivados de áreas como genómica, proteómica, transcriptómica y metabolómica para generar nueva información y conocimientos".

"Si bien existe desde los años 70, recién en el inicio de los 90 se diseñaron e implementaron nuevos algoritmos para el análisis comparativo de secuencias de proteínas y de genes o para la búsqueda de patrones o repeticiones", graficó Rivarola quien planteó que, en el mundo de la bioinformática, este primer gran avance es conocido como el alineamiento de cadenas y de secuencias.

El acceso a las tecnologías de secuenciación de generación avanzada (NGS, por sus siglas en inglés), desde 2007 en adelante, no solo permitió obtener de manera rápida y con gran profundidad el detalle de la secuencia nucleotídica completa de un organismo y comprender su organización, sino que modificó la manera de abordar la genómica.

> "Muchas operaciones informáticas biológicas requieren una gran carga computacional e infraestructura para el almacenamiento de datos debido a la suma y la combinación de información" (M. Rivarola).

"Gracias a estos avances es posible tener una visión completa de un genoma determinado", indicó Rivarola quien añadió: "Esto influyó de manera drástica en programas de mejoramiento genético, aportó mayor competitividad a laboratorios de mediana complejidad y posibilitó el descenso de los costos en la secuenciación de genomas o transcriptomas".

Antes de 2003, fecha en la que se publicó el genoma humano ensamblado, era impensado resolver preguntas vinculadas a cómo enlazar genomas tan grandes. Básicamente, porque era imposible generar los datos y luego procesarlos. "Muchas operaciones informáticas biológicas requieren una gran carga compu-

Ciencia de los datos

Con el objetivo de explorar la diversidad genética, evolución, estructura poblacional, mecanismo de patogenicidad, mecanismos de respuesta a estreses bióticos y abióticos, mapeo genético y mapeo por asociación de caracteres de importancia agrícola y forestal, los equipos del INTA se enfocan en el desarrollo de nuevas metodologías de genómica y bioinformática.

Por su *expertise*, Rivarola aborda el desarrollo y adopción de herramientas bioinformáticas que se aplican al mejoramiento de especies como girasol, soja y sorgo con el foco puesto en la respuesta al estrés biótico –provocado por una plaga o enfermedad–, abiótico –causado por sequía o inundación–, así como procesos fisiológicos que afectan el rendimiento como la senescencia y el brotado precosecha.

Asimismo, la Unidad de Bioinformática del INTA lleva adelante estudios genómicos en especies de interés para la producción de madera (eucaliptos, especies forestales nativas) y frutales (pecan) junto con Susana Marcuchi –especialista del Instituto de Biotecnología del INTA–, Susana Torales –especialista del Instituto de Recursos Biológicos del INTA– e investigadores de las estaciones experimentales Bariloche –Río Negro–, Famaillá –Tucumán–, Montecarlo –Misiones–, Delta –Buenos Aires– y el Instituto de Fisiología y Recursos Genéticos Vegetales (IFRGV), entre otros.

“Apuntamos al desarrollo y utilización de marcadores genéticos útiles para asistir a los programas de mejoramiento en vigencia, aportando a la caracterización de los recursos genéticos, análisis de las poblaciones de mejora y producción, caracterización y conservación de la diversidad”, afirmó Rivarola.

> La Argentina fue el único país latinoamericano que participó del Consorcio de Secuenciación del Genoma del Trigo.

tacional e infraestructura para el almacenamiento de datos debido a la suma y la combinación de información”, manifestó Rivarola.

“En los últimos 15 años, la bioinformática es un campo de investigación que explotó y, sin dudas, es la herramienta para las investigaciones del futuro”, aseguró Rivarola.

Tsunamis de datos

La noticia sobre la secuenciación y ensamblado del genoma del trigo causó una revolución en la comunidad científica y en los medios internacionales. La envergadura y duración del proyecto dejó claro que no se trató de una tarea simple: participaron más de 200 científicos de 73 instituciones, procedentes de 20 países y llevó más de 13 años.

Dirigidos por el Consorcio de Secuenciación del Genoma del Trigo (IWGSC, por sus siglas en inglés), investigadores de todo el mundo presentaron el estudio genético del cereal más detallado hecho hasta el momento. Es como un manual detallado con las instrucciones genéticas que contiene la secuencia del 94 % de los 21 cromosomas. Además, incluye la localización de casi 108.000 de sus genes y la presencia de millones de marcadores y elementos que regulan y controlan los procesos biológicos, que son el resultado del fenotipo.

De la mano del INTA, la Universidad Nacional del Sur (UNS), el Conicet y los servicios genómicos de Indear (empresa público-privada entre Bioceres y Conicet de servicios genómicos), la Argentina fue el único país latinoamericano que participó del IWGSC.

Allí, la Unidad de Bioinformática del INTA junto con Marcelo Helguera, especialista en genética y genómica aplicada al mejoramiento de trigo en el INTA Marco Juárez –Córdoba–, Viviana Echenique, directora del Centro de Recursos Naturales Renovables de la Zona Semiárida (Cerzos) del Conicet en Bahía Blanca –Buenos Aires– y Gabriela Tranquilli, del Instituto de Recursos Biológicos del INTA, participaron en la secuenciación del cromosoma 4D.

“Se trató de un proyecto ambicioso, dado que el genoma del trigo es cono-

cido como uno de los 'gigantes' de las plantas", indicó Helguera y confirmó: "Tiene casi 16 mil millones de pares de bases, esto es el equivalente a cinco veces el genoma humano, 30 veces el de arroz y siete veces el de maíz. Por esto, lograr la secuenciación completa de su genoma representó un gran desafío para la ciencia".

La secuenciación del genoma completo de trigo permitió definir un catálogo de casi 110.000 genes, organizados linealmente por cromosoma. "Esto nos permite referenciar de forma muchísimo más precisa cualquier estudio genético de trigo acelerándose el descubrimiento de genes de interés agronómico y el posterior uso de esta información en los programas de mejoramiento", analizó Helguera.

El trabajo no fue sencillo y el camino recorrido fue largo. Iniciado en 2005, el proyecto del Consorcio Internacional –integrado originalmente por un pequeño grupo de científicos– tuvo dos etapas: en la primera, buscaron obtener la secuencia básica y preliminar de cada cromosoma; en la segunda, obtener una secuencia y ensamblado de alta calidad. "Luego de obtener la secuencia, se realiza un proceso denominado ensamblado, que implica utilizar algoritmos matemáticos para encontrar el orden más preciso de los genes en cada uno de los cromosomas", explicó Tranquilli.

A rigor de verdad, el conocimiento de la secuencia del trigo acelerará la obtención de variedades más resistentes y productivas. "Lo que sigue es empezar a asignar funciones a estos genes, entender cómo se relacionan las redes que forman y, después, diseñar estrategias en el marco de programas de mejoramiento genético", consideró Tranquilli y graficó: "Lo que nos llevaba años descubrir, ahora lo podremos hacer en un corto plazo. Esperamos que el conocimiento y su implementación avancen de forma agigantada".

El equipo argentino tuvo una activa participación durante la primera etapa de este proceso y, gracias a un trabajo de articulación público-privada, los investigadores argentinos se concentraron en la secuenciación y ensamblado del cromosoma 4D.

Para esto, se aplicaron filtros sobre las secuencias crudas obtenidas de la secuenciación del cromosoma 4D para la eliminación de lecturas de baja calidad, programas de ensamblado de las lecturas de alta calidad, procedimientos realizados en el Instituto de Biotecnología del INTA, programas de anotación de genes provenientes del ensamblado previo, en articulación del Instituto de Biotecnología del INTA con el INRA (Francia) y programas de establecimiento de orden virtual de genes, utilizando GenomeZipper desarrollado en el *Munich Information Center for Protein Sequences* (MIPS, por sus siglas en inglés) de Alemania.

De acuerdo con Echenique, el grupo de bioinformáticos del Cerzos colaboró en la anotación de genes y en la búsqueda y clasificación de los elementos repetitivos que abundan en este genoma. "Estos elementos son mencionados en libros viejos de genética como 'ADN basura', dado que no se conocían sus funciones ni su clasificación", señaló.

"Sin embargo, estudios de genómica demostraron que esos elementos repetitivos tienen unas proporciones y ubicaciones características en los distintos genomas, permitiendo inferir que cumplen roles importantes", expresó Echenique quien destacó: "El equipo de investigadores del Cerzos-Conicet tuvo un rol estratégico en el análisis de estos elementos".

"Nuestra participación en este proyecto no solo nos permitió la formación de investigadores en bioinformática, sino que es un avance magnífico desde el punto de vista de soberanía tecnológica", valoró Helguera quien añadió: "Con esta información, podremos acelerar el desarrollo de mapas genéticos de alta definición, el descubrimiento de genes y variantes alélicas superiores".

En la actualidad, programas de mejoramiento genético –mediante el uso de marcadores moleculares– están trabajando en la hibridación y generación de nuevos cultivares.

> "Se trató de un proyecto ambicioso, dado que el genoma del trigo es conocido como uno de los 'gigantes' de las plantas" (Marcelo Helguera).

> "Nuestra participación en este proyecto no solo nos permitió la formación de investigadores en bioinformática, sino que es un avance magnífico desde el punto de vista de soberanía tecnológica" (M. Helguera).

"El nuevo laboratorio de servicio de genotipado de alto caudal –recientemente inaugurado en el Cerzos– acelerará la selección asistida en los programas de mejoramiento públicos y privados de la Argentina", ponderó Echenique.

En este sentido, el nuevo laboratorio cuenta con una plataforma única en el sector público que permite trasladar a los cultivos los logros de la genómica.

El desafío de armar un rompecabezas genético

Secuenciar el genoma del girasol representó un gran reto para la ciencia. Sin embrago, el interés por descifrarlo impulsó a investigadores de todo el mundo a trabajar de manera articulada para trazar el mapa físico de la planta.

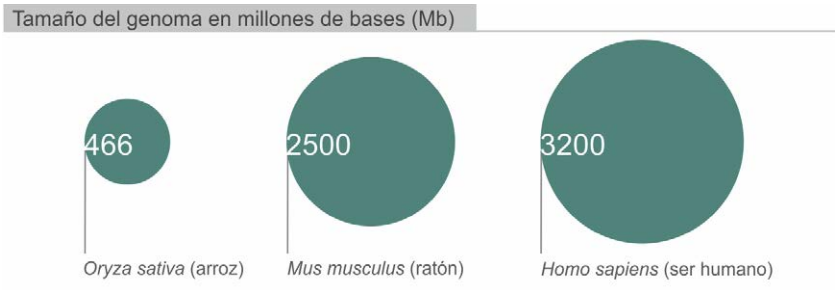
"Es un poco más grande que el genoma humano, está organizado en 17 cromosomas y contiene más de dos tercios de secuencias repetitivas, característica que dificulta su reconstrucción", explicó Norma Paniego, especialista en secuenciación y genotipificación del INTA, y agregó: "Es como armar un rompecabezas con colores similares, lo que dificulta definir la posición de cada pieza".

Con el genoma ensamblado, el siguiente paso será identificar la localización precisa de las regiones del ADN y los genes que definen la capacidad de adaptación al ambiente –resistencia a sequía, frío o suelos salinos–, la resistencia a enfermedades y los rasgos de calidad industrial –aceite o lignina para uso en la producción de energía–.

Está claro que los avances en las tecnologías de secuenciación permitieron mejorar el conocimiento en genómica y entender cómo es y se estructura el ADN. Con el foco puesto en buscar una alternativa que permita superar la complejidad que representan las regiones repetitivas, en 2015 aparecen las tecnologías de secuenciación de tercera generación.

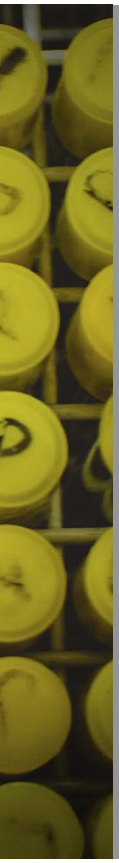
"Este progreso en el campo de la genómica, acompañado por la bioinformática, nos permitió analizar estructuras genómicas complejas, como la del girasol y la del trigo", indicó Paniego quien manifestó que "de esta manera se pudo conseguir un ensamblado bastante preciso del genoma de una línea homocigota francesa de girasol, que es la que ahora se usa como referencia en el mundo".

En general, para la reconstrucción de genomas complejos, se utilizan programas que intentan ensamblar el genoma. Así, mediante la superposición de las distintas lecturas de ADN generadas, que están disponibles en exceso y des-



Comparación de distintos genomas expresados en millones de bases.

Estímulo para la bioinformática en América Latina



Los avances genéticos logrados en los últimos años no hubieran sido posible sin el desarrollo en otras áreas científicas y tecnológicas como electrónica, física y computación, y sus aplicaciones a la biología.

Con el foco puesto en promover el intercambio y el entrenamiento de los investigadores, un consorcio internacional integrado por siete países (Argentina, Brasil, Colombia, Costa Rica, México, Perú y el Reino Unido) da un gran impulso en América Latina a los temas vinculados con la bioinformática. Se trata del Proyecto de fortalecimiento de capacidades para bioinformática (Cabana, por sus siglas en inglés).

"Cabana tiene como objetivo principal capacitar y aumentar la cantidad de investigadores que utilicen herramientas bioinformáticas para acelerar el desarrollo de las ciencias biológicas en América Latina", resumió Rivarola quien aludió a la importancia de fortalecer las redes de investigación existentes en la región sobre este tema.

En este marco, el INTA y el *European Bioinformatic Institute* (EBI) del Reino Unido firmaron un convenio de articulación para capacitar a los profesionales en bioinformática, aplicaciones de genómica, variación genética y transcriptómica.

"Mediante el entrenamiento y la articulación entre las instituciones participantes, este programa pretende aumentar el uso de la bioinformática para proporcionar una formación de alta calidad y explorar una amplia gama de áreas de investigación", indicó el especialista del Conicet y agregó: "La sinergia que logramos con esta propuesta nos permitió generar y compartir datos biológicos en cantidades sin precedentes".

fasadas unas de otras, se consigue reconstruir los cromosomas. Luego, para que el genoma reconstruido se convierta en una referencia se reconocen las estructuras funcionales contenidas en el código, como genes probables, regiones reguladoras y repetitivas, entre otras.

"Cada uno de esos pasos requiere del uso de programas bioinformáticos específicos y algoritmos que favorecen la comparación e integración de datos", expresó Paniego y añadió: "Finalmente, el genoma ensamblado se comparte, desde un portal propio del proyecto del genoma de girasol y también desde repositorios internacionales como el del Instituto Europeo de Bioinformática".

"Identificar cuáles son los genes o las redes genéticas involucradas en el comportamiento de la planta y cuáles son responsables de un determinado rasgo es fundamental para el desarrollo de estrategias de edición génica", explicó Paniego quien destacó el potencial de esa información para acelerar el proceso de mejora sobre los materiales locales adaptados a las distintas regiones de cultivo.

"La Argentina fue pionera en el mejoramiento de girasol, que se inició antes de la creación del INTA", aseguró Paniego para quien el conocimiento de su genoma será de gran ayuda para plantear los procesos de edición de los genes que se quieran modificar.

"Una vez identificados los genes o regiones asociadas a un carácter, disponer del genoma aporta la cartografía exacta de dónde está lo que queremos editar y cuál es su contexto, lo cual es fundamental para el éxito de la estrategia", analizó Paniego.

Con la identificación de los genes que le permiten a la planta resistir frente al ataque de un patógeno, a la falta de agua o la senescencia, se podrá avanzar en el desarrollo de nuevas variedades con mayor calidad y con mejor adaptabilidad.

Más información: *Máximo Rivarola rivarola.maximo@inta.gob.ar; Marcelo Helguera helguera.marcelo@inta.gob.ar; Viviana Echenique echeniq@criba.edu.ar; Gabriela Tranquilli tranquilli.gabriela@inta.gob.ar; Norma Paniego paniego.norma@inta.gob.ar*

"Identificar cuáles son los genes o las redes genéticas involucradas en el comportamiento de la planta y cuáles son responsables de un determinado rasgo es fundamental para el desarrollo de estrategias de edición génica" (Norma Paniego).

Glosario

Genómica, estudio del genoma completo, de todos los genes que se encuentran en un organismo; en contraste, la genética estudia genes de forma individual.

Proteómica, conocimiento a gran escala de las proteínas, en particular de su estructura y función.

Transcriptómica, estudio del conjunto de ARN (ARNr, ARNt, ARNm, ARNi, miARN) que existe en una célula, tejido u órgano.

Metabolómica, estudio de metabolitos o moléculas de bajo peso molecular presentes en las células que dan cuenta del estado funcional de estas.

Hibridación, cruce de dos organismos o de alguna o más cualidades diferentes.

Marcador molecular es un segmento de ADN con una ubicación física identificable (locus) en un cromosoma y cuya herencia genética se puede rastrear.